UNSUPERVISED LEARNING

CLUSTERING

Pendahuluan

- Analisis Cluster adalah salah satu analisis multivariat yang bertujuan untuk mengelompokkan objek-objek berdasarkan kemiripan karakteristik yang dimilikinya.
- Tingkat kemiripan karakteristik objek-objek dalam suatu kelompok (cluster) sangat tinggi, sedangkan tingkat kemiripan karakteristik objek antar cluster satu dengan lainnya rendah.
- Terdapat dua metode cluster yaitu hierarki dan non-hierarki.

Similarity and Dissimilarity Between Objects (review)

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+...+|x_{i_p}-x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

Similarity and Dissimilarity Between Objects (review)

- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- Also, one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

Metode Hierarki

- Metode clustering hierarki dapat dilakukan berdasarkan pendekatan agglomerative (penggabungan / bottom-up) dan divisive (pemisahan / top-down).
- Pendekatan agglomerative menggabungkan satu persatu objek menjadi cluster-cluster baru yang telah ditentukan kedekatan antar clusternya. Proses penentuan kedekataan dilakukan dengan menghitung jarak antar cluster.
- Pendekatan divisive yaitu memulai banyaknya cluster sebanyak satu cluster beranggotakan seluruh objek, kemudian dipisahkan menjadi dua berdasarkan kriteria kedekatan.

Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - ✓Any desired number of clusters can be obtained by ‘cutting’ the dendogram at the proper level

- They may correspond to meaningful taxonomies
  - ✓Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

Metode Hierarki

- Single Linkage (Nearest Neighbor)
- Complete Linkage (Furthest Neighbor)
- Average Linkage (Within Group Linkage)

No	Metode	Jarak antar kelompok (i, j) dengan k
1	Single linkage	$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}$
2	Complete linkage	$d_{(UV)W} = \max\{d_{UW}, d_{VW}\}$
3	Average linkage	$d_{(UV)W} = \text{average}\{d_{UW}, d_{VW}\} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$ d_{ik} : jarak diantara obyek i pada klaster (UV) dan obyek k pada klaster W

Metode Hierarki : Contoh Single Linkage

1

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}$$

Jarak terdekat $\min_{i,k}(d_{ik}) = d_{53} = 2$

5 dan 3 digabung untuk membentuk cluster (35)

Jarak antara cluster (35) dan objek yang lain yang tersisa yaitu 1, :

$d_{(35)1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3$
$d_{(35)2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7$
$d_{(35)4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8$

2

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ \textcircled{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

Jarak terdekat $d_{(35)1} = 3$

35 dan 1 digabung untuk membentuk cluster (35)1

Jarak antara cluster (35)1 dan objek yang lain yang tersisa yaitu 2 dan 4

$d_{(135)2} = \min\{d_{(35)2}, d_{12}\} = \min\{7, 9\} = 7$
$d_{(135)4} = \min\{d_{(35)4}, d_{14}\} = \min\{8, 6\} = 6$

Metode Hierarki : Contoh Single Linkage

3

$$\begin{array}{c} \\ (135) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & \textcircled{5} & 0 \end{bmatrix} \end{array}$$

Jarak terdekat $d_{42} = 5$

4 dan 2 digabung untuk membentuk cluster (24)

Didapatkan 2 cluster yang berlainan, (135) dan (24).

Jarak terdekat adalah

$$d_{(135)24} = \min\{d_{(135)2}, d_{(135)4}\} = \min\{7, 6\} = 6$$

4

$$\begin{array}{c} \\ (135) \\ (24) \end{array} \begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{bmatrix} 0 & \\ \textcircled{6} & 0 \end{bmatrix} \end{array}$$

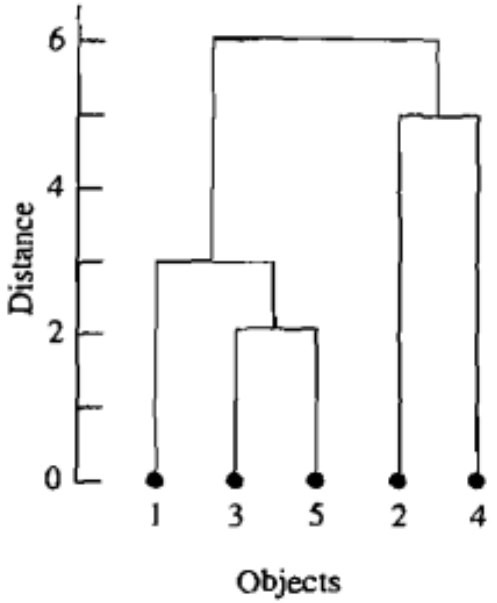
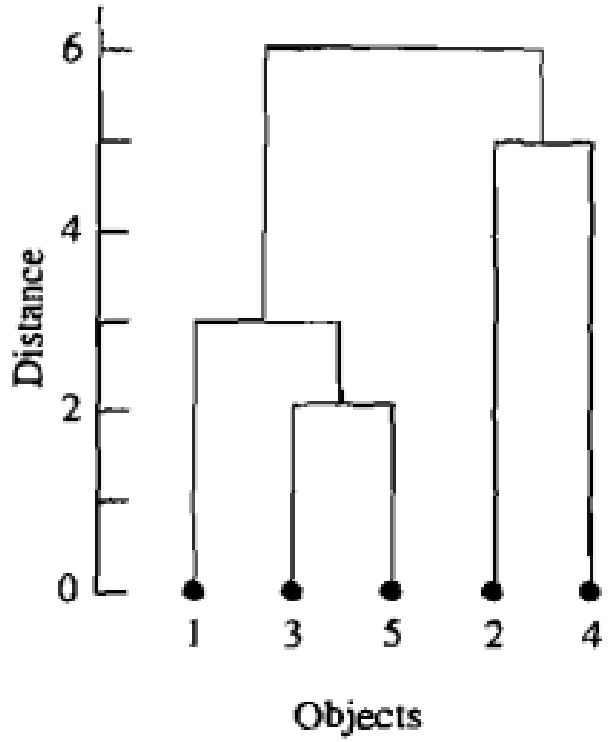


Figure 12.3 Single linkage dendrogram for distances between five objects.

Metode Hierarki : Contoh Single Linkage

Tahap	Jarak penggabungan	Cluster 1 (Anggota)	Cluster 2 (Anggota)	Cluster (Anggota)	Banyak Cluster
0				(1),(2),(3),(4),(5)	5
1	2	(3)	(5)	(3, 5), (1), (2), (4)	4
2	3	(3, 5)	(1)	(1, 3, 5), (2), (4)	3
3	5	(2)	(4)	(1, 3, 5), (2, 4)	2
4	6	(1,3,5)	(2,4)	(1, 3, 5, 2, 4)	1



Strength and Limitation of Single Linkage

- Strength

Can handle non-elliptical shapes

- Limitation

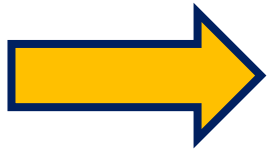
Sensitive to noise and outliers

Metode Hierarki : Contoh Complete Linkage

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}$$

At the first stage, objects 3 and 5 are merged, since they are most similar. This gives the cluster (35). At stage 2, we compute

$$d_{(35)1} = \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11$$
$$d_{(35)2} = \max\{d_{32}, d_{52}\} = 10$$
$$d_{(35)4} = \max\{d_{34}, d_{54}\} = 9$$



$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & \textcircled{5} & 0 \end{bmatrix} \end{array}$$

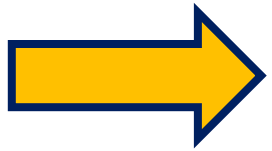
Metode Hierarki : Contoh Complete Linkage

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & \textcircled{5} & 0 \end{bmatrix} \end{array}$$

The next merger occurs between the most similar groups, 2 and 4, to give the cluster (24). At stage 3, we have

$$d_{(24)(35)} = \max\{d_{2(35)}, d_{4(35)}\} = \max\{10, 9\} = 10$$

$$d_{(24)1} = \max\{d_{21}, d_{41}\} = 9$$



$$\begin{array}{c} \\ (35) \\ (24) \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} (35) & (24) & 1 \end{array} \\ \begin{bmatrix} 0 & & \\ 10 & 0 & \\ 11 & \textcircled{9} & 0 \end{bmatrix} \end{array}$$

Metode Hierarki : Contoh Complete Linkage

The next merger produces the cluster (124). At the final stage, the groups (35) and (124) are merged as the single cluster (12345) at level

$$d_{(124)(35)} = \max\{d_{1(35)}, d_{(24)(35)}\} = \max\{11, 10\} = 11$$

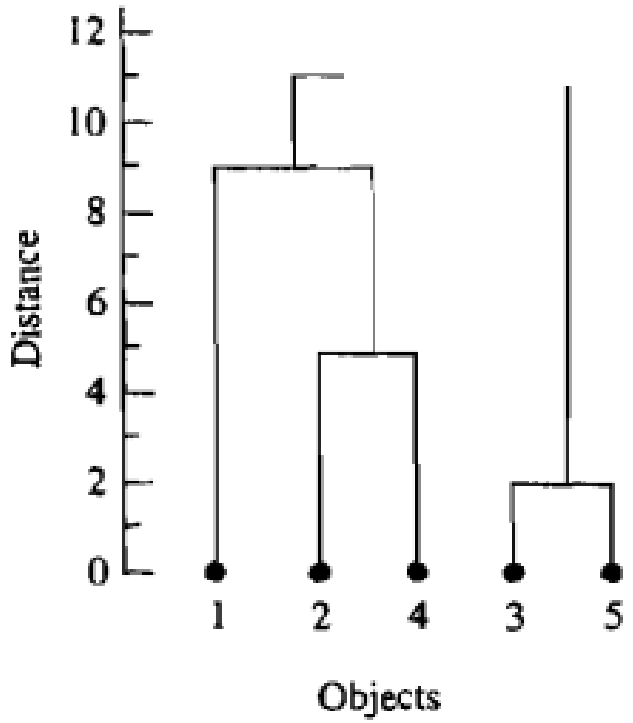
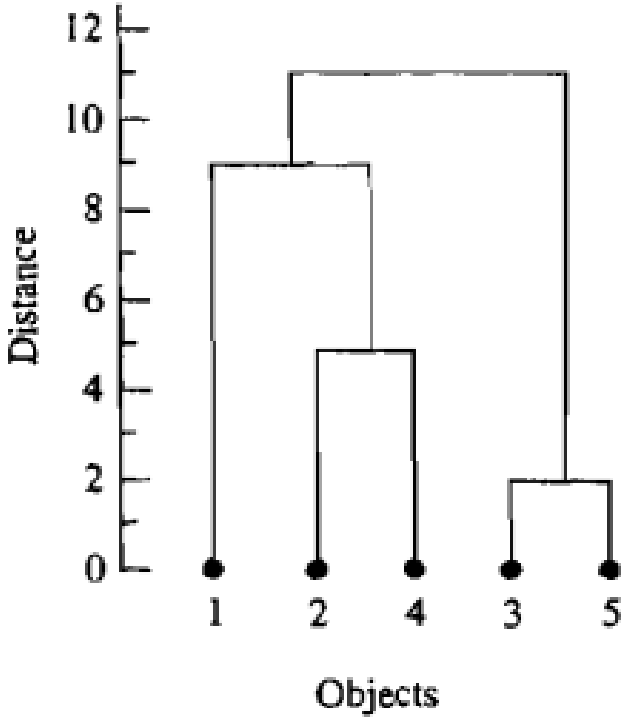


Figure 12.6 Complete linkage dendrogram for distances between five objects.

Metode Hierarki : Contoh Complete Linkage

Tahap	Jarak penggabungan	Cluster 1 (Anggota)	Cluster 2 (Anggota)	Cluster (Anggota)	Banyak Cluster
0				(1),(2),(3),(4),(5)	5
1	2	(3)	(5)	(3, 5), (1), (2), (4)	4
2	5	(2)	(4)	(3, 5), (1), (2), (4)	3
3	9	(2, 4)	(1)	(3, 5), (1, 2, 4)	2
4	11	(3,5)	(1,2,4)	(3, 5, 1, 2, 4)	1



Strength and Limitation of Complete Linkage

- Strength

Less susceptible to noise and outliers

- Limitation
  - ✓ Tends to break large clusters
  - ✓ Biased towards globular clusters

Strength and Limitation of Average Linkage

- Compromise between Single and Complete Link

- Strengths
  - ✓ Less susceptible to noise and outliers

- Limitations
  - ✓ Biased towards globular clusters

Hierarchical Clustering: Problems and Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and convex shapes
  - Breaking large clusters

Metode non-Hierarki

- Metode *K-Means* merupakan salah satu metode analisis klaster non-hirarki yang dapat digunakan untuk mempartisi objek ke dalam kelompok-kelompok berdasarkan kedekatan karakteristik, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu klaster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan ke dalam klaster yang lain.
- Tujuan pengelompokan adalah untuk meminimalkan *objective function* yang di set dalam proses pengelompokan, yang pada dasarnya berusaha untuk meminimalkan variasi dalam satu klaster dan memaksimalkan variasi antar klaster (Johnson, 2007).
- Kelebihan metode K-Means adalah efisien untuk data yang besar.
- Kelemahan metode K-Means adalah jumlah / banyaknya cluster dapat ditentukan di awal oleh peneliti.

Metode non-Hierarki

Algoritma K-Means:

a) Menentukan besarnya *k* (banyaknya cluster yang akan dibentuk) serta centroid awal di tiap cluster. Penentuan centroid awal dapat dilakukan secara acak dari *k* buah observasi.

b) Menghitung jarak antara setiap objek dengan centroid awal, kemudian memasukkan objek-objek ke suatu cluster berdasarkan jarak terdekat dengan centroid yang bersesuaian. Umumnya perhitungan jarak dilakukan berdasarkan jarak euclidean.

c) Menghitung kembali centroid dari *cluster-cluster* yang baru dibentuk

d) Mengulangi langkah (b) dan (c) sampai tidak ada lagi objek yang berpindah cluster

(Johnson, 2007 hal: 696-699)

Metode Non-Hierarki : Contoh K-means

Example 12.11 (Clustering using the *K*-means method) Suppose we measure two variables X_1 and X_2 for each of four items A, B, C, and D. The data are given in the following table:

Item	Observations x_1	x_2
A	5	3
B	−1	1
C	1	−2
D	−3	−2

The objective is to divide these items into $K = 2$ clusters

Metode Non-Hierarki : Contoh K-means

To implement the $K = 2$-means method, we *arbitrarily* partition the items into two clusters, such as (AB) and (CD), and compute the coordinates $(\bar{x}_1, \bar{x}_2)$ of the cluster centroid (mean). Thus, at Step 1, we have

Cluster	Coordinates of centroid	
	$\bar{x}_1$	$\bar{x}_2$
(AB)	$\frac{5 + (-1)}{2} = 2$	$\frac{3 + 1}{2} = 2$
(CD)	$\frac{1 + (-3)}{2} = -1$	$\frac{-2 + (-2)}{2} = -2$

Metode Non-Hierarki : Contoh K-means

At Step 2, we compute the Euclidean distance of each item from the group centroids and reassign each item to the nearest group. If an item is moved from the initial configuration, the cluster centroids (means) must be updated before proceeding. The ith coordinate, $i = 1, 2, \ldots, p$, of the centroid is easily updated using the formulas:

$$\bar{x}_{i,\,new} = \frac{n\bar{x}_i + x_{ji}}{n + 1} \qquad \text{if the jth item is } \textit{added} \text{ to a group}$$

$$\bar{x}_{i,\,new} = \frac{n\bar{x}_i - x_{ji}}{n - 1} \qquad \text{if the jth item is } \textit{removed} \text{ from a group}$$

Here n is the number of items in the "old" group with centroid $\bar{x}' = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$.

Metode Non-Hierarki : Contoh K-means

Consider the initial clusters (AB) and (CD). The **coordinates** of the centroids are $(2, 2)$ and $(-1, -2)$ respectively. Suppose item A with coordinates $(5, 3)$ is moved to the (CD) group. The new groups are (B) and (ACD) with updated centroids:

Group (B) $\quad \bar{x}_{1,\,new} = \dfrac{2(2)-5}{2-1} = -1 \quad \bar{x}_{2,\,new} = \dfrac{2(2)-3}{2-1} = 1$, the **coordinates** of B

Group (ACD) $\bar{x}_{1,\,new} = \dfrac{2(-1)+5}{2+1} = 1 \quad \bar{x}_{2,\,new} = \dfrac{2(-2)+3}{2+1} = -.33$

Metode Non-Hierarki : Contoh K-means

Returning to the initial groupings in Step 1, we compute the squared distances

$$d^2(A,(AB)) = (5-2)^2 + (3-2)^2 = 10$$
$$d^2(A,(CD)) = (5+1)^2 + (3+2)^2 = 61$$

if A is not moved

$$d^2(A,(B)) = (5+1)^2 + (3-1)^2 = 40$$
$$d^2(A,(ACD)) = (5-1)^2 + (3+.33)^2 = 27.09$$

if A is moved to the (CD) group

Since A is closer to the center of (AB) than it is to the center of (ACD), it is not reassigned.

Metode Non-Hierarki : Contoh K-means

Continuing, we consider reassigning B. We get

$$d^2(B,(AB)) = (-1-2)^2 + (1-2)^2 = 10$$
$$d^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9$$

if B is not moved

$$d^2(B,(A)) = (-1-5)^2 + (1-3)^2 = 40$$
$$d^2(B,(BCD)) = (-1+1)^2 + (1+1)^2 = 4$$

if B is moved to the (CD) group

Since B is closer to the center of (BCD) than it is to the center of (AB), B is reassigned to the (CD) group. We now have the clusters (A) and (BCD) with centroid coordinates $(5, 3)$ and $(-1, -1)$ respectively.

Metode Non-Hierarki : Contoh K-means

We check C for reassignment.

$$d^2(C,(A)) = (1-5)^2 + (-2-3)^2 = 41$$
$$d^2(C,(BCD)) = (1+1)^2 + (-2+1)^2 = 5$$

if C is not moved

$$d^2(C,(AC)) = (1-3)^2 + (-2-5)^2 = 10.25$$
$$d^2(C,(BD)) = (1+2)^2 + (-2+5)^2 = 11.25$$

if C is moved to the (A) group

Since C is closer to the center of the BCD group than it is to the center of the AC group, C is not moved. Continuing in this way, we find that no more reassignments

Metode Non-Hierarki : Contoh K-means

For the final clusters, we have

	Squared distances to group centroids			
	Item			
Cluster	*A*	*B*	*C*	*D*
A	0	40	41	89
(*BCD*)	52	4	5	5

The within cluster sum of squares (sum of squared distances to centroid) are

Cluster *A*: 0

Cluster (*BCD*): $4 + 5 + 5 = 14$

Evaluating K-means Clusters

Most common measure is Sum of Squared Error (SSE)

- ✓ For each point, the error is the distance to the nearest cluster
- ✓ To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- ✓ x is a data point in cluster C_i and m_i is the representative point for cluster C_i
  - can show that m_i corresponds to the center (mean) of the cluster
- ✓ Given two clusters, we can choose the one with the smallest error
- ✓ One easy way to reduce SSE is to increase K, the number of clusters
  - a good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Terima Kasih